EXPERIMENTAL IMPLEMENTATION OF PITCH-SYNCHRONOUS SYNTHESIS METHODS FOR THE ROMVOX TEXT-TO-SPEECH SYSTEM

Attila Ferencz[1], Radu Arsinte[1], Istvan Nagy[3], Teodora Ratiu[1],
Maria Ferencz[1], Gavril Toderean[2], Diana Zaiu[2], Tünde-Csilla Kovacs[1], Lajos Simon[1]
[1] Software ITC S.A.,109 Gh. Bilascu Street, 3400, Cluj-Napoca, Romania,
Phone:+40-64-1976681, Fax: +40-64-196787, E-mail: ferencz@utcluj.ro
[2] Technical University of Cluj-Napoca
[3] Music Academy "Gh. Dima" of Cluj-Napoca

***Abstract**: The LPC-MPE synthesis method is an alternative method used for obtaining a better quality of the generated vocal signal, that can be easily implemented in vocal signal coding-decoding systems. Using the method in text-to-speech systems is more difficult because of the modification that must be done on the synthesized vocal signal in order to superimpose prosodical effects. This paper presents our steps in this direction, some researches and experimental results obtained for adapting the system to the pitch-synchronous LPC-MPE method.*

1. INTRODUCTION. PRESENTATION OF THE CLASSICAL ROMVOX SYSTEM

The ROMVOX Text-to-Speech system developed by our team is the first one that allows the synthesis of any unrestricted Romanian text on IBM-PC computers. Diphones are the basic units of the sound inventory. This allows the elimination of the complex transitions' rules between neighbouring phonemes, having a convenient size for the storage space. The first employed synthesis method was the classical LPC (Linear Prediction Coding) with the prediction order 10, the sampling frequency 10 kHz and the frame size 128 or 256 samples. The synthesis algorithm was implemented on a DSPxx25 Development Board based on the TMS320C25 Digital Signal Processor, allowing the real-time synthesis of vocal signals.

The experimental results using the classical LPC synthesis method have proved that the quality of the synthesized signal is limited and it cannot be considerably improved by rising the prediction order, the sampling frequency or the parameters refreshing frequency. This quality limitation is due to the rough approximation of the excitation signal (by periodical Dirac pulse trains or random white noise) and to the voiced/unvoiced decision.

2. THE LPC-MPE SYNTHESIS METHOD

In order to improve the quality of the synthesized signal, several synthesis methods have been experimented, some of them are still under research. The first try in this direction was the experimental implementation of the LPC-MPE (Multi-Pulse Excitation) method. This method does not care about the difference between the voiced and the unvoiced signals. So in the analysis phase an optimally chosen $u(n)$ excitation signal, consisting of a fixed number (M) of excitation pulses, is calculated for each frame.

$$u(n) = \sum_{j=1}^{M} b_j \delta(n - n_j) \qquad (1)$$

The determination of the optimal positions and the corresponding amplitudes for those M pulses is a non-linear problem based on the iterative minimization of the resulted approximation error.

The main idea is to minimize, for each frame of the signal, the energy of the error signal $e(n)$. In this way, it would be possible to determine the optimal values for the parameters b_j and n_j, $j=1,2, \ldots, M$.

The iterative steps to be accomplished in order to determine the amplitude parameters b_j and the position parameters n_j are presented in Figure 1, according to [1] and [2], where $x(n)$ is the initial vocal signal, $\tilde{x}(n)$ is the generated vocal signal and $e(n) = x(n) - \tilde{x}(n)$ is the approximation error.

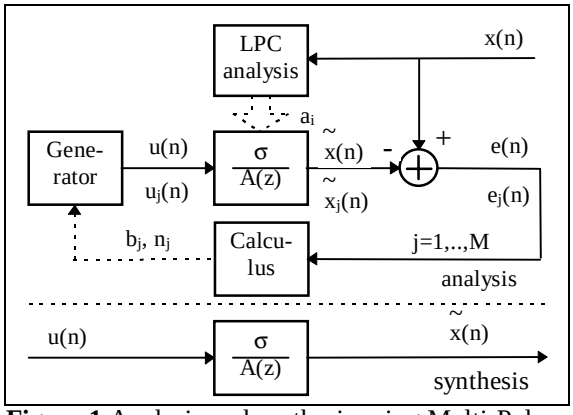

Figure 1 Analysis and synthesis using Multi-Pulse Excitation

The optimal choise of the position parameter n_j could be achieved by looking for the maximum of the term:

$$\frac{\left[\sum_n e_{j-1}(n) h(n - n_j)\right]^2}{\sum_n h^2(n - n_j)} \qquad (2)$$

for n_j taking values along the segment under analysis. The j index signifies the j-th iteration. Here e_{j-1} is the approximation error of the former iteration and h denotes the weight function (the answer to the Dirac pulse) of the filter having the transfer function $\frac{\sigma}{A(z)}$.

After the determination of the optimal value for the n_j position, results the corresponding b_j value. After the iterative determination of the optimal positions n_1, n_2,...,n_M it is possible to globally recompute the amplitudes b_1, b_2,...,b_M following the idea of the global minimization of the energy of the error signal $e_M(n)$. One first step was the implementation of the specific analysis algorithms for the determination of those parameters for each frame of a given vocal signal. Our program allows the graphical visualization of the iterative steps for each frame.

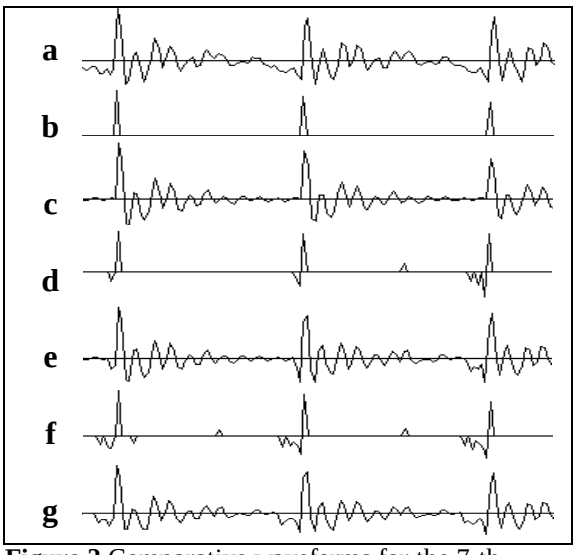

Figure 2 Comparative waveforms for the 7-th frame of the **ae** diphone

a) the initial waveform (pronounced)
b) Dirac pulse train used as excitation signal in LPC classical method
c) the synthesized signal with LPC method
d) the excitation signal determined by LPC-MPE method, M=10
e) the synthesized signal with LPC-MPE method, M=10
f) the excitation signal determined for M=20
g) the synthesized signal for M=20
Note: The **a)** waveform is presented after pre-emphasis and the generated waveforms **c)**, **e)**, **g)** are presented before de-emphasis.

The waveforms presented in Figure 2 provide support for the utility of the LPC-MPE method. We present a frame of the initial waveform of the **ae** voiced diphone chosen from the sound database. It was pronounced by a female. The diphone is divided in frames having 128 samples length. As it results from Figure 2 the excitation waveform **d)** contains, besides the main pulses synchronous with the fundamental frequency, some additional pulses with lower amplitude having the role to correct the distortions mentioned before in the generated waveform **e)**. The result is a better approximation of the initial signal **a).** If we raise the number of excitation pulses to M=20 we obtain a more complex excitation signal **f)**, having more correction pulses around the main excitation pulse and the synthesized signal **g)**, that is even closer to the initial signal **a)**.

The main advantage of this method is a rise of the quality of the synthesized signal and allows a correct synthesis of some special sounds as those with mixed excitation or nasals.

3. EXPERIMENTS REGARDING FUNDAMENTAL FREQUENCY MODIFICATION FOR THE SYNCHRONOUS LPC-MPE SYNTHESIS METHOD

In order to obtain prosodical effects in text-to-speech synthesis systems we have to modify some basic parameters of the synthesized signal as: fundamental frequency, duration or energy. The fundamental frequency could be modified only for voiced sounds, this parameter being undefined for unvoiced sounds, but the LPC-MPE method does not use parameters as fundamental frequency and excitation type. In order to combine the good voice quality obtained by LPC-MPE method with the facility of fundamental frequency modification for voiced signals, we have to detect in the analysis phase parameters like the initial fundamental frequency and the excitation type - as for the classical LPC method - and to handle them in the synthesis phase.

It seems easy to modify the fundamental frequency for obtaining intonation by appropriate modification of the distances between the M excitation pulses for voiced sounds. This modification must be done carefully because, as it can be seen in Figure 2, waveform **f)**, the most pulses are around and very close to the main pulses that determine the beginning of each period. It would be preferable to modify the distances between these groups of pulses rather then the distances inside such a group. Otherwise distortions could appear into the synthesized signal.

We made some experiments that helped us to find a solution as good and simple as possible. The conclusion is presented bellow.

If d_i is the distance between the pulse number i and the preceding pulse, and D_i is the new distance, we have experimentally determined the following relation:

$$D_i = d_i - N\left(1 - \frac{F_{fi}}{F_{fm}}\right)\frac{d_i^2}{\sum_{j=0}^{M} d_j^2} \qquad (3)$$

where N is the frame length, F_{fi} is the initial fundamental frequency and F_{fm} is the modified fundamental frequency. This transformation insures relatively small modifications of the distances between very close pulses and more significant modifications in case of higher distances.

Figure 3 presents the experimental results for two cases of fundamental frequency modification, for an another frame of the same diphone, using M=20. So, the **d)** and **e)** waveforms were obtained for k=1.65, **f)** and **g)** for k=0.75. For the first case the fundamental frequency was raised, for the second one it was lowered. Comparing the waveforms **e)** and **g)**, generated with modified fundamental frequency, with **c)** we can see a right modification of this parameter without any negative influences on the transitory regime inside the period.

On the other side we notice that the useful part of the frames' length is modified proportionally with the fundamental frequency modification. This is an unwanted consequence, the effects of that we have to take into account when we are dealing with rhythm aspects (to make some corrections). Taking into account that the lengthening corrections must be done in the framework of an integer number of the pitch-periods (for voiced sounds) it seems necessary to detect in the analysis phase of the sound inventory the initial positions of these frames. So we have modified our analysis algorithm to permit pitch-synchronous framing for voiced sounds.

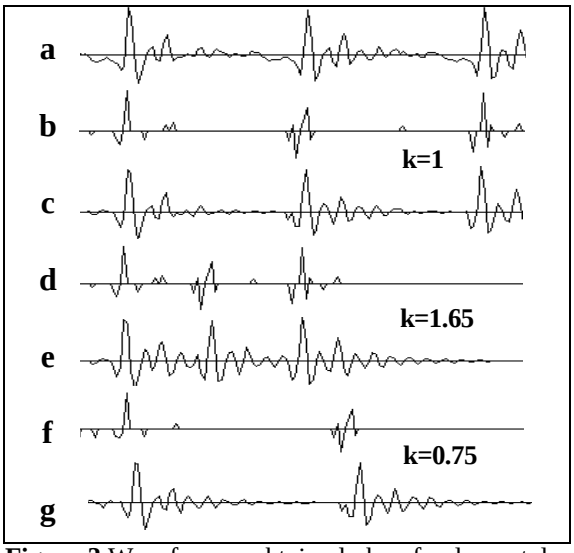

Figure 3 Waveforms obtained by fundamental frequency modification for the 5-th frame of the **ae** diphone, *M*=20

a) initial waveform (pronounced)
b) initial excitation signal
c) synthesized signal (no fundamental frequency modification)
d) modified excitation signal for k=1.65
e) synthesized signal obtained for k=1.65
f) modified excitation signal for k=0.75
g) synthesized signal obtained for k=0.75

4. EXPERIMENTS REGARDING PITCH-SYNCHRONOUS FRAMING OF THE SOUND INVENTORY

Considering the various methods developed for pitch-synchronous framing we adopted on the basis of our experiments a method based also on the LPC-MPE method. The main idea of this method derives from the observation that in the iterative steps described in chapter 2 for voiced sounds in general the first iteration detects an excitation pulse localized at the beginning of a new pitch-period, the second iteration detects one localized at the beginning of another pitch period, a.s.o. (supposing that the initial asynchronous frame contains a few numbers of pitch periods). Also it seems that these excitation pulses localized at the beginning of a pitch-period have in general the highest amplitude from the ***M*** specified. So the modified analysis algorithm consists of two important phases.

- In the first phase an asynchronous framing and analysis procedure is performed. The main purpose of this phase is to localize the voiced and unvoiced zones of the signal and to localize the positions of each pitch-period for the voiced sounds.

- The second phase resumes the analysis on the base of the results of the first step. So in this phase the framing is performed in concordance with the positions detected before.

To improve the performance of the first step the LPC-MPE analysis method was modified in order not to allow detection of multiple excitation pulses situated too close to one another (in the neighbourhood of the main pulses, belonging to the same group described in the previous chapter). The iterative steps described are presented in Figure 4 where the five pitch-periods are detected in six steps (waveform **b)** corresponding to initial signal **a)**). Other possible excitation pulses detected (with smaller amplitudes) are ignored. The iterations according to the next asynchronous frame are made taking into account the results and the initial conditions of the previous frame.

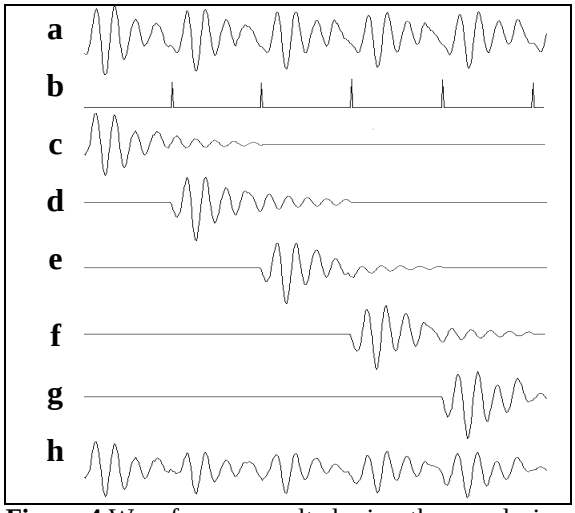

Figure 4 Waveforms resulted in the analysis phase

a) initial waveform (pronounced)
b) pitch-synchronous main excitation impulses
c)-g) frames number 1-5 resulted in the second phase of analysis
h) synthesized waveform as the result of superimposing of the frames number 1-5

In the second phase the LPC-MPE analysis algorithm performs in the normal manner described in chapter 2, but in concordance with the frame sizes and positions detected in the first phase. The result of the analysis from a frame belonging to the second phase is presented in Figure 4, waveforms **c)** to **g)**.

During synthesis the modification of the fundamental frequency can be done by positioning the initial frames at a distance equal with the desired period. Thus, for diminishing the fundamental frequency, the frames, initially having common boundary, are moved away from one another, proportionally with the desired modification of the fundamental frequency. To avoid the

lengthening of the whole sequence, after a given number of frames one frame was omitted, the number depending on the ratio of the desired fundamental frequency and the initial one. The omission of some frames in the above mentioned manner is possible due to the insignificant difference of two neighboured, pitch synchronous frames of the periodical signals. An example for this case is presented in Figure 5 where the fundamental frequency was diminished to k=0.66. For maintaining the initial length of the sequence, frames number 1, 2, 4, 5 were used, frame number 3 was omitted.

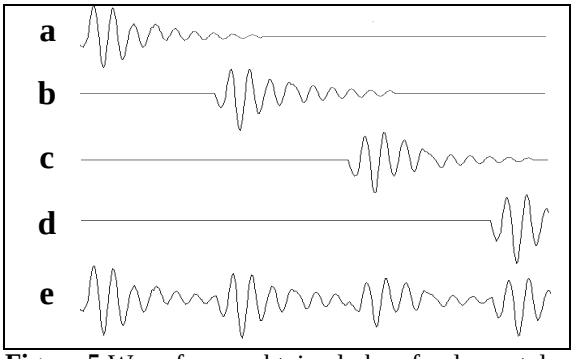

Figure 5 Waweforms obtained by fundamental frequency modification for k=0.66

a)-d) frames number 1, 2, 4, 5
e) synthesized waveform as the result of superimposing of the given frames

To increase the frequency it is necessary to reduce the period of the signal, that is equivalent with the superposition of a part of two consecutive frames, this superposition being proportional with the desired fundamental frequency modification. For maintaining the initial length of the whole sequence, some of these frames must be repeated resulting the necessary increase of the whole duration until the initial one. An example for this case is presented in Figure 6 where the fundamental frequency was increased to k=1.6. For maintaining the initial length of the sequence, frames number 1, 3, 5 were used twice, resulting the sequence with frames number 1, 1, 2, 3, 3, 4, 5, 5.

For obtaining the desired prosodical effects the following must be taken also into consideration:

A. The fundamental frequency can be modified as mentioned above, being the main element of intonation.

B. The rhythm (the speed of the speech) can be increased for voiced sounds by the periodic omission of some frames, depending on the desired modification, and the reduction of this speed can be done by periodical repetition of some frames, omission or repetition taking into account the lengthening modification due to the modification of the fundamental frequency.

In the case of the unvoiced sounds (without any periodicity) the lengthening of the frames is done by the repetition of the final part of each frame, with the length proportional with the desired length modification. The shortening of the frames is made also by the simple omission of a final fragment of these frames.

C. Besides the lengthening of the accentuated syllable, the accent also requires a signal with higher energy, respectively the excitation pulses must have a higher amplitude.

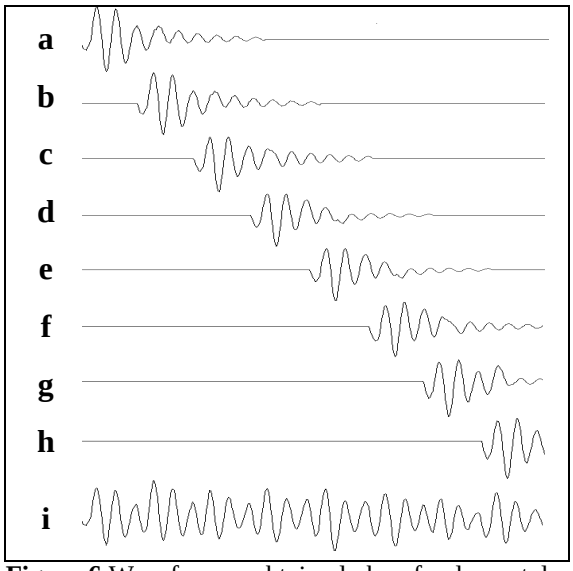

Figure 6 Waveforms obtained by fundamental frequency modification for k=1.6

a)-h) frames number 1, 2, 3, 4, 5
i) synthesized waveform as the result of superimposing of the frames number 1, 1, 2, 3, 3, 4, 5, 5

5. CONCLUSIONS CONCERNING THE LPC-MPE SYNTHESIS METHOD

The experiments we have done prove that the quality of the generated signal is better than that synthesized by the classical LPC method. This could be explained by the optimally chosen positions of the excitation pulses supplied to the synthesis filter.

The signal quality of the two presented LPC-MPE (asynchronous and pitch-synchronous) methods are the same but the first one presents a few inconveniences regarding prosodical effects superimposing. So the only usable method in text-to-speech synthesis is the second one.

The disadvantage of the method is the lower processing speed in the analysis phase, due to the algorithm's iterations. Also, the modification of the fundamental frequency in the synthesis phase is more difficult to be achieved than in the case of the classical LPC method.

6. REFERENCES

[1] B. J. Atal, M. R. Schroeder, "*A new model of LPC excitation for producing natural-sounding speech at low- bit rates*", Proc. IEEE ICASSP, 1982
[2] Rene Boite, Murat Kunt, "*Traitement de la parole*", Press Politechniques Romandes, 1987
[3] G. Olaszy, G. Németh, "*Multilingual Text-to-Speech Converter*", Journal on Communications, No. 2, 1991
[4] A. Ferencz, a. o., "*Experimental Implementation of the LPC-MPE Synthesis Method for the ROMVOX Text-to-Speech System*", SPECOM '96, St. Petersburg, Russia, October 1996